

Bibliometric Networks

Frank Havemann^{*} and Andrea Scharnhorst^{**}

^{*}Institut für Bibliotheks- und Informationswissenschaft der Humboldt-Universität zu Berlin

^{**}Data Archiving and Networked Services (DANS), Den Haag

Abstract

This text is based on a translation of a chapter in a handbook about network analysis (published in German) where we tried to make beginners familiar with some basic notions and recent developments of network analysis applied to bibliometric issues (Havemann and Scharnhorst 2010). We have added some recent references.

Contents

1	Introduction	1
2	Citation Networks of Articles	2
3	Bibliographic Coupling	5
4	Co-citation Networks	6
5	Citation Networks of Journals	8
5.1	Citation Flows Between Journals	9
5.2	Citation Environments of Individual Journals	10
6	Lexical Coupling and Co-Word Analysis	10
7	Co-authorship Networks	11
8	Outlook	14

1 Introduction

Bibliometrics is a research field that deals with the statistical analysis of bibliographic information. Most bibliometric research can be categorized as scientometric, because it relates to scientific publication output, in particular to journal articles. Informetrics captures flows of information not just within, but also beyond the world of books and periodicals, including communication over the Web (Webometrics) and the Internet.¹

The bibliographic description of a written work contains a number of elements such as the name(s) of the author(s), title of the piece, keywords and data necessary for locating the document (e.g., title of the journal or edited volume in which an article appears, year of publication, volume/issue number, page numbers). This information is often collected and archived in databases. All of these elements constitute the bibliographic attributes of a document, also called the metadata.

A document's attributes are connected to one another through the document itself—author(s) to journal, keywords to publication date, etc. These connections of different attributes gener-

¹For an introduction to these overlapping fields, we refer the reader to the *Introduction to Informetrics*, a textbook by Egghe and Rousseau (1990), to the *Handbook of Quantitative Science and Technology Research*, edited by Moed, Glänzel, and Schmoch (2004), to *Bibliometrics and Citation Analysis*, the most recent English language monograph by De Bellis (2009), and to the recent open-access electronic book by Havemann (2009), *Einführung in die Bibliometrie* (Introduction to Bibliometrics, some of the sections in this article are essentially abbreviated versions of sections contained in chapter three of Havemann's book.) Finally, as a good starting point to initiate the subject, we refer the reader as well to the lecture text of Wolfgang Glänzel (2003), *Bibliometrics as a Research Field*.

ate bipartite networks which can be represented as rectangular matrices. Like attributes, such as authors with authors (see section 7 below, on co-authorship networks) or keywords with keywords (see section 6 below, on co-word analysis), can also be coupled to one another; this kind of coupling produces unipartite networks, represented by square, symmetrical matrices.

Scientific publications are characterized by the fact that, as a rule, they contain references to other scientific works. This generates a further network, namely, that of the publications themselves. This kind of self-reference has also been taken up in models of the scientific publication process (Bruckner et al. 1990; Gilbert 1997; Leydesdorff 2001; Morris et al. 2003; Lucio-Arias and Leydesdorff 2009).

After Eugene Garfield’s pioneering and seminal work, the *Science Citation Index* (SCI), in the 1960s, the SCI was used not only to navigate information, but it was also applied as a tool in scientific analysis (Garfield 1955; de Solla Price 1965; Wouters 1999). The growing accessibility of machine-readable data (the SCI was already available on magnet tapes relatively early on) and the appearance of the second big information network, the *World Wide Web* (WWW), made large-scale automated data processing and analysis possible. The WWW—in which pages constitute the network nodes and (hyper-)links the edges—has itself become a popular object of research in network analysis (Huberman 2001).

Link analysis also proves fruitful when applied to academic institutions or countries, as was shown by Thelwall (2004, 2009) and also by Ortega et al. (2008). The position of links in a network, allows us to draw conclusions about research collaboration and the function of different national scientific systems in international research landscape.

In addition, the Web remains today a medium for accessing data, in part freely, which we can analyze according to their own particular network structures. Consider, for example, the bio-information databases, Ebay, or Facebook. Data-flow networks of this kind have resulted in the evolution of a new specialty within statistical physics, namely, that of *complex networks* (Scharnhorst 2003; Morris and Yen 2004; Pyka and Scharnhorst 2009). Within the multidisciplinary setting of network science, the methods of *social network analy-*

sis (SNA), originally developed for smaller social networks, are combined with statistical analysis and dynamic modeling from physics, with computer science algorithms for data mining and visualization, and with graph theory in mathematics, for the purpose of better grasping, explaining, and mastering existing complex networks in nature and society (NRC 2005). In recent years, bibliometrics and scientometrics have been strongly influenced by these developments in new network research (Börner et al. 2007). We will examine this more closely in section 8 below. In what follows immediately, however, we confine ourselves to the network descriptions that have traditionally received special attention in scientometrics.

In scientometrics we distinguish roughly between the analysis of texts and the analysis of actors. Even if information on both elements is contained in a single bibliographic record, historically speaking, most bibliometric network analyses have concentrated on text elements (viz., citation, co-citation, bibliographic coupling, or semantic networks). This stands in contrast to the analysis of scientific collaboration networks—either for cooperation on an individual level or cooperation between countries (Wagner 2008). Among the more interesting methods that have evolved in the intermediate field of text and actor is the HistCite method developed by Eugene Garfield et al., which we discuss in section 2 below.

2 Citation Networks of Articles

Articles in scientific periodicals base on other knowledge and acknowledge this by referencing earlier articles and other publications. This way they build networks. This view of a stream of periodicals literature was already propagated more than 40 years ago by Derek J. de Solla Price (1965). Defining articles as nodes or vertices and citations as the links or edges of a network graph allows us to apply graph theoretical methods to bibliometrics. In view of the social nature of the science system, we will also consider how terms and concepts initially developed for SNA can be fruitfully applied to explain scientific communication.

New nodes with directed edges pointing to previously available nodes are continually added to both the network of journal articles and the Web. Whereas web pages can be modified (or even completely deleted), a journal article remains an unchanged document once it has been published. Corrections, for example, can only be added to the document subsequently as *errata* or *corrigenda*. Hyperlinks can be added to existing pages retroactively, referring to pages that have been subsequently produced. In citation networks, by contrast, there is temporal ordering; but the order, as it were, is not strict. Because of the protractedness of the publication process, authors may be aware of not-yet-published works and cite these, but citation network analysis usually does not take this into account.

Today’s vast network of scientific journal articles began to build up when the referencing of older publications became common practice. It might be helpful to try to visualize the spacial expansion of the whole article citation network as a continually growing sphere which adds a new “growth ring” every year, in which articles are located through the sources that they cite. The HistCite method developed by Eugene Garfield et al. (2003) assumes that, at the root of ev-

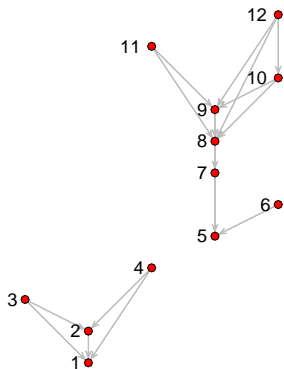


Figure 1: Temporally ordered graph of a citation network for the first twelve articles on N-rays. Data source: de Solla Price (1965)

ery citation network, there is a single piece of path-breaking research—namely, the major work of one scientist or the core works emanating from a specialist group or specialty area. Successful strands of research can thus be extracted or distilled from this work. In easily accessible networks of frequently cited articles, the paths of scientific insight and knowledge are clearly visible. This method can also be used to show connections between different scientific schools and communities or their relative isolation from each other (Lucio-Arias and Scharnhorst 2012).

As Lucio-Arias and Leydesdorff (2008) show, a main path analysis of these temporally ordered graphs reveals the mechanisms of dissemination and diversification (diffusion), as well as those of consolidation and standardization (codification) of scientific knowledge. This kind of citation-based historiography complements biographic and scientific-historic investigation and, in so doing, bridges the graph-theoretical concept of networks and the role of social networks in structuralist social theory (Merton 1957).

If we determine the actual location of a publication only on the basis of information given in the sources cited, then essentially we forgo other information in the document, which may be equally crucial for this determination. Therefore, a reconstruction of knowledge transfer should not buttress itself solely on the analysis of citation networks. Citation analysis does, however, have the advantage of being able to use the mathematical calculations of SNA and thereby achieves results that hermeneutic historiography alone cannot. Moreover the standard methods of information retrieval use the textual similarities between documents to show users of specifically retrieved texts further texts that might be relevant for them (see section 6, S. 10). More recent approaches to information retrieval embrace also other bibliometric regularities (Mutschke et al. 2011), co-author patterns, and journal distributions.

Relations (edges, links) in a network of nodes can be captured mathematically in a square, adjacency matrix A whose elements $a_{ij} \geq 0$ are different from zero, if node i is related to node j . If we do not differentiate between relations of different strengths, then a_{ij} has a value of 1, if a relationship exists.

In his above-mentioned article, Derek J. de Solla Price (1965) presented the adjacency

matrix for citations between articles in a self-contained bibliography on N-rays, whereby the ones were symbolized with dots and the places that would have been filled by the zeros were left blank (de Solla Price 1965, p. 514: figure 6).² He ordered the articles in temporal sequence according to their date of publication and omitted citations of all sources external to the bibliography. The adjacency matrix for the first twelve articles is thus

$$A = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 0 & 0 \end{pmatrix}.$$

Because future articles could not be cited, ones occur in rows of this matrix only to the left of (or below) the main diagonal. Thus, because of the temporal sequencing of the citation network, de Solla Price's adjacency matrix takes the form of a triangle; along the main diagonal and to the right of (or above) it occur only zeros ($a_{ij} = 0, \forall j \geq i$).³

In figure 1, the graph of the first twelve articles in the citation network disaggregate into two partial graphs. Each of these graphs represents independently achieved results which could only, at a subsequent point in time—namely, with the appearance of the first overview article on N-rays (number 75 in the bibliography)—be interpreted as belonging together.⁴ The adjacency matrix, A , of a citation network can be used to model a reader's behavior moving from article to article by following the cited sources. For example, a reader may begin with article 12; the starting time thus noted is $t = 0$. He/she can be described by the column vector, $\vec{r}(0)$, which contains eleven zeros and one one as the twelfth component. By multiplying this vector from the left with the transpose

of A , one finds the reader by articles 8, 9, and 10. In the next step, by means of the rule $\vec{r}' \leftarrow A^T \vec{r}$, he/she wanders from there to articles 7, 8, and 9, and so forth:

$$\vec{r}(0) = \begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 1 \end{pmatrix}, \vec{r}(1) = \begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 1 \\ 1 \\ 1 \\ 0 \\ 0 \\ 0 \end{pmatrix}, \vec{r}(2) = \begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 1 \\ 2 \\ 1 \\ 0 \\ 0 \\ 0 \end{pmatrix}, \dots$$

We can refine this model to one of the random reader,⁵ whereby each component of the vector scales to one: $\vec{R} = \vec{r} / \sum r_i \equiv \vec{r} / r_+$. Thus, for a given time $t = 2$, for example, the components of the reader vector different from zero will be $R_7(2) = 1/4$, $R_8(2) = 1/2$, $R_9(2) = 1/4$. By scaling to one, we can interpret these fractions as a probability, namely, the probability that, at time $t = 2$ we would encounter the reader by article 7, 8, or 9, should he/she, upon completion of his/her reading, randomly select one of the sources cited. We have a double chance of finding the reader by article 8 because he/she can arrive at 8 via 9 as well as 10. This makes it clear why scaling to one allows us now to designate the reader as a random reader.

Today, with the aid of citation indexes, readers are not only able to search for articles in the reference lists of cited sources retrospectively, but they can also navigate temporally forward in citation networks. We can model this process by using the transpose of the transposed adjacency matrix that is, A itself, $(A^T)^T = A$, because reflection along the main diagonals reverses all of the arrows in the graph of a directed network, which is easy to show.

The adjacency matrix, A , gives the direct paths between the nodes in the network along the directed edges. For the model of the reader, we used powers of A (or A^T): A^1, A^2, A^3, \dots . For example, if we compare A^2 with the graph in figure 1, we see that the components of A^2 indicate how many (indirect) paths of length 2 there are between the nodes—for instance, there are two such

²N-rays turned out to be fictive; so, for that reason, the bibliography qualifies as self-contained. This concrete example of a citation graph will accompany us through the subsequent sections of this paper.

³Temporally ordered networks are acyclic. Within acyclic graphs, there is no path along the directed links which loops back to return to the starting point.

⁴This expresses itself as co-citation (cf. section 4, p. 6).

⁵This refers to the random surfer model upon which Brin and Page (1998) based their PageRank algorithm.

paths between node 8 and node 12 (between the remaining pairs of nodes there is at most one path of length 2). In general, then, a matrix A^k contains the number of ways of length k between the vertexes.

3 Bibliographic Coupling

In accordance with Kessler (1963), two articles are said to be bibliographically coupled if at least one cited source appears in the bibliographies or reference lists of both articles. If we look for such bibliographic couplings in the graph of the first twelve articles on N-rays (figure 1, p. 3), we discover a number of these. Nodes 2 and 3 are bibliographically coupled over node 1, as are nodes 2 and 4. Nodes 3 and 4 are coupled over node 1 and over node 2. Nodes 6 and 7 are coupled over node 5. Nodes 9 through 12 are coupled in pairs over node 8, and nodes 10 to 12 are coupled in pairs additionally over node 9.

If a reader desires to locate a bibliographically coupled article in a citation network, he/she must first move one step along an arrow to the next vertex, and then, from that node, in the opposite direction of the next arrow. As known from graph theory, this process is described by the matrix $B = AA^T$. Its element b_{ij} , in accordance with the rules for matrix multiplication, is the scalar product of the row vectors from A . Because A is a binary matrix, summation results in the number of matching components of both row vectors, that is, the number of common sources.

Thus, element b_{ij} indicates how many bibliographic couplings exist between articles i and j . In other words, b_{ij} gives the number of paths of length 2, via which one moves from i along the arrow and then to j in the opposite direction. The symmetry of the coupled pairs corresponds to that of the matrix $B = B^T$. The main diagonal contains the numbers of bibliographic self-couplings of an article, namely, the numbers of all of their references to other articles in the network.

Citation databases enable users to move temporally forward, backward, and laterally (by zig-zagging). Thematically similar articles appearing in the same volume of a journal are often temporally so proximate that the earlier article cannot be cited in the later one. But they also reveal their likeness through similar reference lists, that is, through strong bibliographic coupling. As early

as the late-1980s, with the old CD-ROM edition of the Science Citation Index, the user was directed from an article which he/she located to the twenty most strongly bibliographically coupled articles via the "related records" option.⁶ The strength of the coupling of two articles, i and j is defined here simply by the number of references that the articles have in common, as given by the element b_{ij} of matrix B .

In our example of the bibliography on N-rays, we can only include in our analysis the citation relations between articles in the N-ray bibliography, although clearly other sources have been cited in the articles' reference lists, which do not belong to the bibliography. In an alternative approach, we can analyze a complete body of scientific literature of a publication period together with all cited sources. Using the SCI we could, for instance, analyze the publications in one specific year. Rather than selecting a thematic excerpt or segment from the citation graph, we then consider all of the journal articles of that year with all of their cited sources, including books, patents, newspaper articles, etc. Accordingly, the resultant matrix is not a square adjacency matrix A whose rows and columns represent the same vertexes, but rather a rectangular matrix, each of whose rows represents an article and each of whose columns represents a cited source. Only a few of the journal articles for that particular year will appear as a source. Thus we arrive at citation network consisting of just two types of nodes: articles and sources. Note that the articles carry the same publication year, and the sources can be from any year. Within this network, only edges between nodes of different type are permitted. Networks of this type are also called *bipartite*; the rectangular matrix is termed an affiliation matrix.⁷

For the rectangular affiliation matrix A with m rows (articles) and n columns (sources), we can also calculate the matrix of bibliographic couplings $B = AA^T$. Matrix B is also square, in this case, and contains for each of the m articles one row and one column.

Two articles, both with long reference lists, could contain many sources in common, in which case they would be said to be strongly bibliograph-

⁶This use of bibliographic coupling for information retrieval is still part of the on-line edition of the Science Citation Index, now part of the Web of Knowledge (see <http://wokinfo.com/>).

⁷From the Latin *ad-filiare* meaning to adopt as a son.

ically coupled. Articles with only a few references, therefore, would tend to be more weakly bibliographically coupled, if coupling strength is measured simply according to the number of references articles contain in common. This suggests that it might be more practicable to switch to a relative measure of bibliographic coupling, which we can define most simply by using set theory. In references lists, each cited source appears only once; thus we can see a reference list as a set of cited sources. In set theory language we formulate this thus:

$$b_{ij} = |\mathbf{R}_i \cap \mathbf{R}_j|,$$

that is, the element b_{ij} of matrix B equals the size of the intersection of the reference lists in articles i and j . The Jaccard index (or Jaccard similarity coefficient) gives us a relative measure of the overlap of two sets:⁸

$$J_{ij} = \frac{|\mathbf{R}_i \cap \mathbf{R}_j|}{|\mathbf{R}_i \cup \mathbf{R}_j|}. \quad (1)$$

The Jaccard index of bibliographic coupling is zero if the intersection of the reference lists is empty; it reaches a maximum of one if both lists are identical (because, in this case, the intersection would be equal to the union).

Another relative measure of similarity between sets, which we can use here, is the so-called Salton index:⁹

$$S_{ij} = \frac{|\mathbf{R}_i \cap \mathbf{R}_j|}{\sqrt{|\mathbf{R}_i| |\mathbf{R}_j|}}. \quad (2)$$

In this case, the average size of the sets is related to the geometric mean of the size of both sets. Here, too, the index reaches a maximum of one for identical sets and a minimum of zero for disjoint ones.

4 Co-citation Networks

We speak of the co-citation of two articles when both are cited in a third article. Thus, co-citation

can be seen as the counterpart of bibliographic coupling. Returning to our example, we can also find a number of instances of co-citation among the first twelve articles on N-rays (figure 1): articles 1 and 2 are co-cited twice (in articles 3 and 4); articles 8 and 9 are co-cited three times (in articles 10, 11, and 12, respectively), and article 10 is co-cited once with article 8 and once with article 9 (in article 12).

In the previous section we explained how the bibliographic coupling matrix B can be obtained from the scalar product of the row vectors of the adjacency matrix A . Now, we have to construct the scalar product of the column vectors from A , in order to calculate the elements for the co-citation matrix C :

$$c_{ij} = \sum_k a_{ki} a_{kj}.$$

Since A is a binary matrix, the summation yields the number of common components of the respective column vectors, that is, the number of cases in which articles appear in the same row or reference list. Written compactly, we calculate $C = A^T A$. Thus our model reader moves within the graph, first, in the opposite direction of the arrow and then, in a second step, with the arrow. Like matrix B , matrix C is also symmetric. The main diagonal of C contains the number of cases in which an article is co-cited with itself (which is the case for every citation). The number c_{ii} is therefore the number of all citations of article i in other articles in the network.

Most of the elements in the co-citation matrix C of our example are equal to zero. This is so because the content-related connections between both citation strands only became apparent as co-citation of these papers, initially, with the publication of the first overview article on N-rays (number 75 in the N-rays bibliography and thus not visible in figure 1). Co-citation relations, unlike bibliographic couplings, are subject to change. Many content-related connections can or may only be recognized by later authors at some subsequent point in time; or, conversely, can or may be deemed as no longer essential by later authors.

As with the bibliographic coupling of articles, it is equally reasonable for purposes of co-citation analysis to consider all of the journal articles of a year with all of their cited sources. We will now analyze the bipartite network of articles and

⁸The Swiss botanist and plant physiologist, Paul Jaccard (1868–1944), defined this index in 1901.

⁹The computer scientist Gerard Salton (1927–1995), who lived and worked in the United States, was one of the pioneers in the area of information retrieval (cf. Wikipedia). This index was introduced in Salton and McGill (1983). In section 3.6 of the above-cited bibliometrics textbook, Havemann (2009) shows how Salton and McGill define their index alternatively as the cosine of the angle between the row vectors of matrix A .

sources introduced in the previous section not according to how the articles are coupled over sources, but rather just the opposite: that is, how co-citation in articles connects the sources to one another. We can also calculate matrix C in accordance with the bipartite network, yielding $C = A^T A$.

In the co-citation analysis of two successive volumes of a bibliographic database, many co-cited sources in the first year's papers will appear again as co-cited sources in the second. Coupling strength will vary, however; many new co-cited pairs of sources will join the old ones. The reference lists for a given year's papers, practically speaking, are analogous to the results of an opinion survey designed to ascertain which cited sources are currently seen to be related to one another.

The principle of co-citation was first applied by Irina Marshakova (1973) in Moscow in a study on laser physics. Independently from Marshakova the principle was also propagated by Henry Small (1973), a scientist from the Institute for Scientific Information (ISI) in Philadelphia, founded by Eugene Garfield in 1960. Since the 1970s, the co-citation perspective has been at the core of bibliometric analyses of specialties, fields, scientific schools, or paradigms in the sense of Kuhn—in other words, co-citation has provided the main thrust underlying our understanding of the social and cognitive theoretical structure of science (De Bellis 2009).

All bibliometric networks can be visualized or modeled graphically. (We will come back to this in section 8 below.) Historically, co-citation analysis data were used for the mapping of science and for the first Atlas of Science project (Garfield 1981). For such purposes, the network of co-cited sources is calculated or derived from the reference lists of publications from a single year's papers, whereby only those sources are considered the citation numbers of which exceed a certain threshold value (e.g., a threshold value of 5). These sources were seen by Henry G. Small (1978) as *concept symbols*. In a network thus constructed, the next step is to try to determine clusters of sources, whose members have a strong relationship to each other, but relate only weakly to sources in other clusters.¹⁰

¹⁰In network analysis, such clusters of nodes are also called *communities*.

To determine clusters of similar objects in accordance with the above criteria, a set of algorithms was developed. If we wish to apply these algorithms, we first have to decide which measure of co-citation strength between two sources best applies: the absolute number of co-citations or, for instance, one of the two relative measures presented above—the Jaccard or Salton indexes.

Relative measures of co-citation result in a weak coupling strength among frequently cited sources, which are only rarely co-cited. This seems appropriate because many of the citing authors do not see a closer relationship between those cited *concept symbols*. At the ISI, Henry Small first started working with the Jaccard index and later with the Salton index (Small and Sweeney 1985). Irina Marshakova (1973) did not use a relative measure. Instead she calculated the expected values for co-citation numbers, based upon the independence of both citation processes, and accepted only those numbers that exceeded the expected values significantly.

In order to get from clusters of *concept symbols* to a map, the next step is aggregation. Hereby, we take all of the nodes in one cluster and draw them together into a point. Then, we take all of the links between each set of two clusters and draw these together into a single link of a determined strength that corresponds to the specific factual distance between those clusters. In this way, we create a network of clusters that can be visualized. The technique of multidimensional scaling or MDS has been frequently used to visualize complex networks on a two-dimensional plane. Today, networks are often visualized using force directed placement or FDP.

Since the clusters of nodes thus produced, in turn, represent the vertexes in a co-citation network; using an analogous clustering procedure, we can now generate clusters of clusters and so on, until all of the cited sources of a given year's papers are united in a single cluster that represents that part of science indexed by the database used.

Co-citation cannot only be applied on articles. We can also inquire, for example, how often authors are cited together in order to model or map the structure of the expert community. Or, we can analyze co-citations of entire journals (see section 5). In neither case, however, can we expect that aggregations of papers represent just one specific subject. Authors,

for instance, usually deal with several themes—sometimes simultaneously—which can also belong to different specialized areas of expertise. Another problem of author co-citation is the growing tendency toward more research collaboration, which becomes visible in the increasing number of authors per article in some fields. The thematic flexibility of authors leads in fact to a real problem: If we wish to determine or define authors’ areas of activity, thematic flexibility turns out to be a crucial factor whenever we seek to trace dynamic processes in science (on this matter, see section 8 below).

5 Citation Networks of Journals

Research has been able to expand so much over the centuries only because new areas of expertise have continually opened up, and the various research tasks have been distributed accordingly among the expert communities representing these new fields of scientific endeavor. Each community has created its own specialized journal. Alongside to these specialized journals, journals coexist in which research results of general interest are published. Currently, the open-access movement changes the journal landscape profoundly.¹¹ Still, we can assume that the foundation of a journal is a response of a communicative need of a scientific community in one field, one country or across several.

But even the most highly specialized journals contain more than just those articles contributed by the experts in their respective fields of research: these periodicals also contain articles from other research areas that could be of interest for any reader of a particular journal at a given time. The result of this is that the literature from one scientific area is not just to be found in the core journals of that area, but rather that it is broadly scattered according to Bradford’s law (Bradford 1934).

Nevertheless, despite the addition of literature external to a core research field and in accordance with the Porphyrian tree of knowledge, articles in one journal should cite the articles of journals from contiguous fields more often than they would those from fields or disciplines further away. The

early study by Gross and Gross (1927) was based on this plausible assumption, already. They determined the number of citations of other journals in the general chemistry publication *Journal of the American Chemical Society*—for the purpose of providing librarians with data relevant for library journal selection. However, number-of-journal-citations data gathered in this way are not just influenced by thematic proximity or distance, but also simply by the quantity and quality of the articles in the journal referenced. Journals with similar topics compete for the articles that are most important for further research. For that reason, articles submitted to a journal for publication must undergo a qualitative appraisal process, the so-called *peer review*. The bigger a journal’s reputation, the more articles it will be offered, and hence the more rigorous its review process is likely to be. Thus scientific periodicals differ not only according to area of expertise but also according to reputation.

In sum, then, the citation flows in a network of scientific journals are influenced by three main factors: thematic contiguity, the size of a journal, and the reputation of a journal. As a further influencing variable can be added the usual number of cited sources per article for the particular area of specialty in question.

We can correct for journal size by relating the number of citations to the corresponding number of articles available to be cited, as Garfield and Sher (1963) did when they introduced the *journal impact factor* (JIF). In order to take account of the different citation behavior customary in different fields, Pinski and Narin (1976) suggested that instead of relating the number of journal citations to the number of citable articles, this number should be related to the total number of references in all of the cited journal’s articles. This is tantamount to a kind of import-export relationship that would also take into account that review articles with long reference lists are on average cited more frequently than original articles publishing research results. Thus journal citation networks, constructed with such a normalization, will just mirror the actual thematic relationships of those periodicals and their respective reputations.

Compared to article citation networks, journals networks will naturally have substantially fewer nodes and are, for that reason, not only more transparent, but also lend themselves more easily

¹¹see the Public Knowledge Project <http://pkp.sfu.ca/>

to numeric analysis. The citation numbers necessary to construct a journals network are presented in aggregated form in the *Journal Citation Reports* of the Science Citation Index and the Social Sciences Citation Index. In the following, we present two examples of journal network analyses.

5.1 Citation Flows Between Journals

First, we will examine different variants of a network consisting of five information science journals: (1) *Information Processing and Management*, (2) *Journal of the American Society for Information Science and Technology*, (3) *Journal of Documentation*, (4) *Journal of Information Science*, and (5) *Scientometrics*.

We begin by constructing a network that has been weighted with reciprocal citation numbers (including self-citations by the journals in the network). With data from the Social Sciences Edition of the Journal Citation Reports, we derive the following adjacency matrix for the citation window 2006 and the publication window 2002–2006:

$$A = \begin{pmatrix} 79 & 65 & 15 & 6 & 24 \\ 42 & 182 & 11 & 15 & 44 \\ 6 & 22 & 37 & 8 & 6 \\ 20 & 26 & 13 & 30 & 11 \\ 7 & 48 & 7 & 10 & 254 \end{pmatrix}. \quad (3)$$

Matrix A contains only elements $a_{ij} \neq 0$, because all five journals cite each other as well as themselves. The main diagonal contains the journal self-citations. In the graph of A , edges (links) between all the five vertices flow in both directions. Loops represent self-citations.

Let us, now, like Pinski and Narin (1976) switch to a network where the number of citations a_{ij} of journal j by journal i are divided by the sum a_{j+} of the number of references to j in the 5-journals network, yielding $\gamma_{ij} = a_{ij}/a_{j+}$.¹²

Pinski and Narin go even further. They argue that citation in a prestigious journal should count more than citation in a less important one. But precisely because they measure prestige itself by number of citations (per cited source), they end up with a recursive concept of that notion, like the concept of prestige that has been debated since the 1940s for social network analysis (Wasserman

and Faust 1994). In social networking, it is not only important how many people one knows; one must know the “right” people, namely, those who know a lot of others who are the “right” ones.

How do we conceptualize this recursive notion of prestige mathematically? To this end, we will examine a model of prestige redistribution for the five information science journals under consideration here. To begin ($t = 0$), all five journals should have the same weight, so we fix this at 1. The weights are written as a column vector. Analogous to our procedure for the reader model in an article citation network, we then multiply the column vector from the left with the transpose of the adjacency matrix $\vec{w}(1) = \gamma^T \vec{w}(0)$. In this way, weights are redistributed within the network. The new column vector contains exactly the row sums of γ^T , i.e. $w_j(1) = \gamma_{+j} = a_{+j}/a_{j+}$, the import-export ratios of journals. By repeating this procedure many times over, we can see that the weights iteratively approach fixed limit values. In accordance with this, for $t \rightarrow \infty$, the following equation holds:¹³

$$\vec{w} = \gamma^T \vec{w}.$$

This means that the weights determined by the iteration fulfill an equation which can be seen as an expression of the recursive definition of prestige, viz., that the weight of journal j results from the citation relations to all of the other journals (as well as to itself), according to how close these relations are, factored by the weight of the other journals. Pinski and Narin call this *influence weight*. Determination equations of this type are also referred to as *bootstrap relations*.

Important to note here is that the iteration procedure is somewhat incorrectly labeled “redistribution.” The sum of the five weights after the first iteration step is $4.76 < n = 5$. In other words, weight is lost (because the import-export relations of the larger journals are more propitious than those of the smaller ones). However, this discrepancy can be corrected through scaling; for $t \rightarrow \infty$, we obtain the normalized components 0.76, 1.33, 1.03, 0.64, and 1.25. By scaling to n , it becomes patently clear who the winners and losers of redistribution are.

¹³That this relationship holds not just in our special case, but in every case, is guaranteed by the theory of eigenvalues. Matrices of type γ^T have a principal (maximal) eigenvalue of 1, and the iteration determines the corresponding eigenvector.

¹²The actual data can be found in the book by Havemann (2009).

Following [Pinski and Narin](#), Nancy [Geller \(1978\)](#) considered a kind of modified redistribution of journal weights. Her starting point was the theory of Markov chains, a special class of stochastic or random processes in which (just as it is for our case here) the situation at time $t + 1$ is completely determined by the situation at time t . Instead of using γ , she took a somewhat differently scaled matrix γ^* with the elements $\gamma_{ij}^* = a_{ij}/a_{i+}$, whose transpose belongs to the stochastic matrices. Stochastic matrices have the property that they leave the sum of vector elements unchanged. Such matrices describe true redistribution; they are therefore well-suited for the description of random processes in which probability is redistributed over possible system states. Nancy [Geller's](#) algorithm is also interesting because there are only a few steps that separate it from Google's *PageRank* algorithm as presented in the textbook by [Havemann \(2009, section 3.3\)](#). This is an example for a method developed for citation networks which became useful also for Web analysis.

5.2 Citation Environments of Individual Journals

If we consider citation matrices for large groups of journals, we see that many cells in such a matrix are empty and that citation tends to be confined to smaller more densely networked groups ([Leydesdorff 2007](#)). This is not surprising; it mirrors the real-world design and relations of scientific specialty areas and disciplines. Nevertheless, the delineation of areas remains a problem which has not been satisfactorily resolved: the borders are fluid.

[Leydesdorff \(2007\)](#) suggested another method for determining the position of individual journals within the mass of and relative to all of the areas of science. The starting point is an ego network; let us take, for example, the journal *Social Networks*. *Social Networks* has two citation environments. The first consists of those journals, in a specific time frame, that cite articles in *Social Networks* from a unique time period. We could call this group or set of journals the awareness area, spillover area, or influence area of *Social Networks*—in other words, its *citation impact environment*. The second environment consists of those journals that are cited in articles in *Social*

Networks—that is, its *knowledge base*. For each group of journals, then, it is possible to carry out the following analysis independently. We ascertain all of the reciprocal citation links for each respective group with the aid of the Journal Citation Reports. *Social Networks*, our starting point, is a member of both environments. For these citation environments we obtain asymmetric matrices like in equation 3 (p. 9).

By applying the process described above, we obtain groups of journals that have similar citation behaviors; these groups can thus be interpreted or defined as specialty areas. The position of the journal whose ego network was at the starting point gives us information about aspects of interdisciplinarity (for example, by applying betweenness centrality) and a possible interface function ([Leydesdorff 2007](#)). If we repeat this analysis over a sequence of years, the sometimes changing function of a journal in an equally dynamic journal environment becomes visible. In the case of *Social Networks*, what was revealed was that this journal's functioning as a possible bridge between traditional areas of social science and new methods and approaches from network theory in physics could be reduced to a single year, namely, 2004 ([Leydesdorff and Schank 2008](#)).

6 Lexical Coupling and Co-Word Analysis

For computer-supported information retrieval (IR), documents are characterized by the terms used in them. A manageable (not too big) but nevertheless informative set of such terms, comprised mainly of keywords supplied by authors or indexers, or significant words in a document's title, can serve well for IR. In the extreme, all of the words in a document can be taken into account. What is interesting then is the frequency with which these words occur, not counting stop words like "the," "and," "of," and others. The appearance of terms in a collection or set of documents is described by the term-document matrix A , whose element a_{ij} tells us how often in document i the term j occurs.

The term-document matrix, like the matrix introduced above consisting of documents and cited sources (see section 3), describes a bipartite network, namely, one of terms and documents. In

this case, however, by taking into account the frequency with which terms occur in a document, we weight the network.

The so-called vector-space model of IR not only reveals the similarity between documents based on the terms used in them (this is lexical coupling), it also shows the relationships between the terms based on their common usage in the documents (this is the basis of co-word analysis). The former corresponds to bibliographic coupling of articles; the latter refers to the co-citation of sources (section 4).

At the beginning of the 1980s Michel Callon and his group in France proposed co-word analyses as an alternative or supplement to prevailing co-citation methods (Callon, Courtial, Turner, and Bauin 1983). So-called “cognitive scientometrics” is supposed to make the relationships between documents clearly evident where, for whatever reason, reciprocal citation has occurred only sparsely. Anthony van Raan’s group in the Netherlands developed interactive tools for co-word based science maps in the context of evaluations. These maps can make the activities of institutions or countries in specific fields of science visible (Noyons 2004). Clusters in these networks were interpreted as themes or topics, and temporally hierarchically branched trees provided some insight into the dynamics of scientific areas (Rip and Courtial 1984).

Co-word analyses can be understood as the empirical method of the so-called actor-network theory, a social theory in the field of science and technology studies (Latour 2005; De Bellis 2009). Despite more recent and promising text-based network analyses for identifying innovation-relevant scientific research—some researchers even speak of literature-based discoveries (Swanson 1986; Kostoff 2007)—expert knowledge for the interpretation of text-based agglomerations is still necessary. And, despite decades-long efforts by many groups, as Howard White put it succinctly in a 2007 discussion, we are still not able to analyze and visualize the development and change of scientific paradigms or scientific controversies in such a way that they are accessible to non-experts.¹⁴

¹⁴Howard White, 11th International Conference of Scientometrics and Informetrics, Madrid 2007, workshop on mapping, personal notes.

This deficit may be due in part to the ambivalence of language. In a study by Leydesdorff and Hellsten (2006), the authors pointed to the significance of context for words, and they suggested returning to the text-document matrices for word analyses as well, in order to gain more complete information. Probably the answer also lies in the clever selection of a base unit for statistical procedures. An analysis of developing discussion focal points in online forums has shown that, already in one single post, contributors brought up or referred to different topics. Therefore, in this particular case, the choice of sentences as the base unit for statistical network analyses produced better results than did the longer text passages of one post (Prabowo, Thelwall, Hellsten, and Scharnhorst 2008).

In text mining, in sources (such as all keywords, all titles, abstracts, or full text) extraction of terms or phrases is performed according to different algorithms; and statistical measures for frequency and correlation (including network indicators) are applied for which the reference unit such as a phrase, a sentence, a document is an important parameter. The plethora of combinations of these elements presents a great challenge to text analysis and text mining—a challenge which can only be addressed through strong networking of all text-based structural searches including semantic Web research (van der Eijk, van Mulligen, Kors, Mons, and van den Berg 2004).

One method of information retrieval developed for extracting topics from corpora, which uses both types of links in bipartite networks of documents and terms—namely, co-word analysis and lexical coupling—is latent semantic analysis (LSA) proposed by Deerwester, Dumais, Furnas, Landauer, and Harshman (1990). This method is based on singular value decomposition (SVD) of the term-document matrix. By means of SVD, bipartite networks of articles and cited sources can also be analyzed thematically (Janssens, Glänzel, and De Moor 2008; Mitesser, Heinz, Havemann, and Gläser 2008).

7 Co-authorship Networks

Co-authorship is considered an indicator of cooperation. If two or more authors share the responsibility for a publication presenting particular research results; then, in the course of the research

process that led to those results, these individuals ought to have worked together in some way or another at one time.

It is appropriate and important to agree on a concept of cooperation before the structure of co-authorship networks is analyzed or interpreted (Sonnenwald 2007). To discuss this in-depth, however, would exceed the scope of this article, so we refer the reader to the definition of research collaboration in the aforementioned bibliometrics textbook (Havemann 2009, section 3.7) where the author relies on work of Grit Laudel (1999, S. 32–40), see also Laudel (2002).

Not every form of collaboration finds its ultimate expression in a co-authored work. On the other hand, a certain tendency to arbitrarily assign co-authorship to individuals (or force one on them) cannot be denied. On the other hand, shared responsibility for an article in a renowned journal is only rarely possible without some form of cooperation. The co-authors, one would expect, at least know each other,¹⁵ and this realistic expectation is what makes the analysis of co-authorship networks interesting.

Co-authorship networks are usually introduced as networks of authors. In its simplest form the co-authorship network is unweighted. A link between two authors exists, if both appear together as authors of at least one publication in the bibliography or body of literature under investigation. Weighting the link with the number of articles in which both appear together as authors suggests itself, but this kind of modeling still uses only part of the information about cooperation that can actually be extracted from a bibliography. What it fails to capture is whether the relationships between authors are purely bilateral or whether these researchers work together in larger groups. If, for example, three authors cooperate on one article, then between them in total three links of weight 1 can be found; this is exactly the same as if three pairs of them had each published one article together (in total, then, three articles).

Another aspect which could be taken into account is the sequence in which the co-authors are listed. Though, in different communities there are different rules whom to place first and last, it has been proposed to include information on author sequence in bibliometric indicators (Galam

2011). More complete information is used if co-authorship is presented as a bipartite network of authors and articles. Element a_{ij} of affiliation matrix A is equal to 1 if author i appears among the authors listed for article j ; otherwise it is equal to zero.

Up to now, most investigations have confined themselves to co-authorship networks in which only authors are represented as nodes and publications are not. The adjacency matrix B of such a network can be calculated from the affiliation matrix A , whereby $B = AA^T$. This becomes immediately apparent if we carry our deliberations on networks of bibliographically coupled articles (section 3) over to the authors-articles network. In a bipartite network, a co-author is someone whom we reach, whenever we take a step toward a publication (A^T) and then back again to an author (A). We derive the co-authorship figures for two authors from the scalar product of the (binary) row vectors of A . The diagonal element b_{ii} in matrix B is therefore equal to the number of publications to which author i was a contributor.

So, how are co-authorship networks structured? First of all, we frequently find that an overwhelming majority of specialty area authors (more than 80 %) are grouped together in a single component of the network, namely, the so-called main component. The remaining authors, conversely, often comprise only small groups of researchers connected to one another (at least indirectly) over co-authorship links. All of the distances occurring between the cooperation partners—whether these are functional (subject-related), institutional, geographical, language-related, cultural, or political—cannot prevent the emergence of one big interrelated network of cooperating scientists.

Equally worthy of note, in comparison to size, are the very slight distances between authors in the main components of co-authorship networks. In a co-authorship network consisting of more than one million biomedical science authors whose combined output for the period 1995–99 was more than two million published articles (verified in Medline), the statistical physicist, Mark E. J. Newman (2001a) found that over 90 % of the authors were grouped together in the main component.¹⁶ The second largest component of this

¹⁵The likely exception here being articles with 100 or more authors.

¹⁶Because authors in bibliographic databases are not always clearly identifiable, the figures vary according to the method of identification used. Newman found that, if he

network contained only 49 authors. The maximal distance between two authors in the main component (also called the network diameter) is a matter of only 24 steps or hops: that is, on the shortest path between two arbitrary nodes, lie a maximum of 23 other nodes. The average length of all of the shortest paths between nodes in the main component is less than five hops (Newman 2001b):

This “small world” effect, first described by Milgram,¹⁷ is, like the existence of a giant component,¹⁸ probably a good sign for science; it shows that scientific information—discoveries, experimental results, theories—will not have far to travel through the network of scientific acquaintance to reach the ears of those who can benefit by them. (Newman 2001b, p. 3)

Newman’s statement restricts itself to the informal communication of results between researchers, which so often precedes formal publication. This observation confirms one of the two behavioral principles of science Manfred Bonitz (1991) formulated early in the history of scientometrics. They read as:

Holographic principle: Scientific information “so behaves” that it is eventually stored everywhere. Scientists “so behave” that they gain access to their information from everywhere.

Maximum speed principle: Scientific information “so behaves” that it reaches its destination in the shortest possible time. Scientists “so behave” that they acquire their information in the shortest possible time.

Newman’s observation confirms the second of Bonitz’ principles.

The famous “small world” experiment referred to above, undertaken by social psychologist Stanley Milgram (1967) for the acquaintance network

took into account authors’ full names, the main component would comprise 1.5 million different authors; however, if he included the surnames and only initials for first name, then this figure shrank to just under 1.1 million individuals. For statistical analysis purposes, this ambiguity is only of minor importance; but it strongly impairs the systematic pursuit of individual research. Newer methods depend on a combination of names, addresses, channels of publication, and citation behavior, in order to automatically and unambiguously ascribe researcher identification.

¹⁷Milgram (1967)

¹⁸Newman (2001a)

in the US, resulted in an average distance of six hops.¹⁹ Newman explains the small-world effect taking himself as an example: he has 26 co-authors who, in turn, author publications together with a total of 623 other researchers.

The “radius” of the whole network around me is reached when the number of neighbors within that radius equals the number of scientists in the giant component of the network, and if the increase in numbers of neighbors with distance continues at the impressive rate [...] it will not take many steps to reach this point. (Newman 2001b, p. 3)

The number of co-authors that an author has is a measure of his/her interconnectedness. It is equal to the number of his/her links (degree of the node) in the co-authorship network. In a given specialty area, marginal or peripheral authors have only a few cooperation partners. In network analysis, therefore, the degree of a node is also a measure of its centrality. Often the distribution of co-authorship is skewed: a few authors have many co-authors, many authors have only a few co-authors (Newman 2001a, p. 5).

Another measure of centrality is the betweenness of a given node i . Betweenness is defined as the total number of shortest paths between arbitrary pairs of nodes, which run through node i . Conceivable then, is that nodes with higher values of betweenness are responsible for shorter distances in networks and also for the emergence of large main components. And, in terms of betweenness centrality, a number of frontrunners also set themselves clearly apart from the remaining authors (Newman 2001b, p. 2).

Finally, the different roles and functions of researchers in co-authorship networks is a topic that has begun to receive increasing attention. Lambiotte and Panzarasa (2009) have recently considered the importance of a researcher’s function (viz., having a high level of connectivity and authority within a community versus being an facilitator or communicator between communities) for the dissemination of new ideas.

¹⁹Milgram’s subjects had the task of sending a letter, via their acquaintances, as close as possible to a recipient unknown to themselves. The letters that actually reached the targeted individual had been forwarded on average by six persons (including the subject).

8 Outlook

Paul Otlet, the pioneer in knowledge organization inherited us a wealth of drawings of the evolving universe of knowledge (van den Heuvel and Rayward 2011). Among them, he depicted the main classes of his decimal classification scheme²⁰ as longitudes of a globe of knowledge (van den Heuvel and Rayward 2005; van den Heuvel 2008). Katy Börner has inspired an artistic visualization of the sciences as a growing “organism” (Börner and Scharnhorst 2009). Recent efforts for a geography or geology of the expanding globe of scientific literature, have led to a revival of “science maps.” The Atlas of Science, presented by the ISI in the early 1980s (Garfield 1981) has been followed by a New Atlas of Science (Börner 2010).²¹ Nevertheless scientists are still struggling to find adequate imagery or, as the case may be, an appropriate mathematical model to represent the deeper understanding we have of the dynamic processes of evolving science networks.

Against this backdrop of the “new network science”, the future of bibliometric network analyses lies in the incorporation of methods and tools from other disciplinary areas. Visualizations represent a possible platform for interdisciplinary encounters (Börner 2010). These new “maps of science” have met with similar controversy to that which we know from the history of geographical maps. It therefore comes as no surprise that mappers of science have sought to build bridges to cartography. If we apply the methods of structure identification (self-organized maps) as they were developed for complexity research (Agarwal and Skupin 2008), then it is just a small step from the question of possible models to the explication of complex structure formation.

²⁰Universal Decimal Classification UDC, see also <http://udcc.org/>

²¹The Atlas of Science of Katy Börner captures parts of a remarkable long-term project, called “Places and Spaces”, which is a growing exhibition of science maps. This exhibition is particularly interesting, for one, because the public invitation and selection process enables highly varied and unique depictions to achieve greater visibility through public display and, second, because the themes of the yearly iterations embrace such a broad and diverse spectrum of subject matter—running, for example, from the history of cartography, over maps as deceptive or phantasy pictures, up to maps drawn by children. Many of the maps on display are based on network data. For more information, see <http://www.scimaps.org>.

For bibliometric networks, next to the traditional topic of structure identity, the topic of structure formation—and with it, the dimension of time—steps more forcefully into the foreground of interest. All of the methods we have used up to now in the global cartography of science or knowledge landscapes (Scharnhorst 2001) confirm the existence of collectively generated self-organizing structures in the form of scientific disciplines and scientific communities. On the macro-level, these patterns are so persistent that, as Klavans and Boyack (2009) have recently shown, they manifest themselves recurrently, relatively independent of the respective research methods.²²

It is more difficult to find general patterns or regularities on the micro-level for the interactions between authors or for those between authors and documented knowledge. What remains is a deeper understanding of the dynamic mechanisms that describe the emergence of a science landscape and individual navigation within it. We expect dynamic mathematical models to reproduce already known statistical bibliometric laws like Lotka’s law of scientific productivity (Lotka 1926) or Bradford’s law of scattering (Bradford 1934), mentioned above. Complexity research with notions like energy, entropy, or fitness landscapes (Scharnhorst 2001) offers a rich repertoire of contemporary analytical methods and models which can do justice to the network character of complex systems (Fronczak, Fronczak, and Hołyst 2007). Recent empirical research in this area is devoted to contemporary effects in evolving bibliometric networks (Börner, Maru, and Goldstone 2004) and the search for burst phenomena (Chen, Chen, Horowitz, Hou, Liu, and Pellegrino 2009) or the application of epidemic models to the dissemination of ideas (Bettencourt, Kaiser, and Kaur 2009). But also on the level of conceptual models philosophy and sociology of science have embraced the idea of an epistemic landscape and mathematical models for the search behavior of researcher in it (Payette 2012; Edmonds, Gilbert, Ahrweiler, and Scharnhorst 2011).

Topic delineation on both the micro as the macro level is a pertinent problem of science studies in general and bibliometrics in particular. However, the problem of topic delineation

²²The ring structure of the consensus map bears astounding similarity to Otlet’s historical visions of a globe of knowledge.

in citation-based networks of papers might be a consequence of the overlap of thematic structures. The overlap of themes in publications is well known to science studies. A recent study by Havemann, Gläser, Heinz, and Struck (2012) on the meso level tests three local approaches to the identification of overlapping communities developed in the last years (Fortunato 2010).

The heuristic value of dynamic models lies in their broad range of available possible mechanisms, forms of interaction, and feedback, which can be connected to distinctive types of structure formation. Thus, for example, the topically relevant notion of field mobility²³ can be drawn upon for the explication of growth processes in competing scientific areas, for which, in addition, schools of knowledge as sources of self-accelerating growth are also highly relevant (Bruckner and Scharnhorst 1986). Through mobility, a network is created between scientific areas, and, at the same time, all of the elementary dynamic model mechanisms and a quasi “metabolic network” of knowledge systems are formed (Ebeling and Scharnhorst 2009). The wanderings of the researcher can be followed or read from his/her self-citations. Whereas self-referencing is usually ignored in scientometrics, these citations constitute an excellent source for the investigation and analysis of mobility in scientific fields. Structures in the self-citation network can be interpreted as thematic areas which become visible through different groups of keywords and co-authorships. The wandering of an individual between research areas, as shown by his/her lifework of collected scientific output, can be displayed or represented as a unique bar code pattern (Hellsten, Lambiotte, Scharnhorst, and Ausloos 2007). This creative fuzziness of the scientist can also be shown on science maps as a spreading phenomenon, whereby the scientist, instead of the wandering dot, is depicted as a flow field (Skupin 2009).

The use of models as generators of hypotheses presupposes that the hypotheses have been empirically tested and, accordingly, put into the context of social communication, socioeconomic, and political theories of “science qua social sys-

tem” (Gläser 2006). This connection to traditionally strongly sociologically anchored SNA—especially the proposed inclusion of social, cognitive, and personality attributes of actors for modeling the evolution of networks and dissemination phenomena within networks—and the inclusion/application of dynamic modeling approaches in SNA provide an excellent basic framework for using models to generate theory (Snijders, van de Bunt, and Steglich 2010).

Semantic web applications creating new linked repositories of data, publication and concepts; large scale data mining techniques (as originating from Artificial Intelligence), meaningful visualizations, and mathematical models of science all contribute to a better understanding of the science system. However, similar to the variety of possible network visualizations is the variety of mathematical and conceptual models of science. Often knowledge about data mining, modeling and visualizing is inherited by different relative isolated academic tribes (Becher and Trowler 2001). Translating and linking concepts and methods where appropriate and possible is one remaining task. Exploring and better understanding our own science history is another one. Only if we succeed in bridging unique individual science biographies with the laws and regularities of the science system as a whole, will we be able to learn more about the development of new ideas—knowledge generation—and be better able to intervene in this process in a more controlled and supportive manner.

Acknowledgement

We thank Mary Elizabeth Kelley-Bibra for helping us with the translation of the text into English.

References

- Agarwal, P. and A. Skupin (2008). *Self-organising maps: applications in geographic information science*. Wiley-Blackwell.
- Becher, T. and P. R. Trowler (2001). *Academic Tribes and Territories: Intellectual enquiry and the culture of disciplines* (2nd ed.). Society for Research into Higher Education / Open University Press.

²³The term “field mobility” was introduced in bibliometrics by Jan Vlachý (1978) to describe the thematic wandering of scientists. Thematic wandering results from new discoveries as well as other grounds such as the connectivity between scientific areas (Bruckner, Ebeling, and Scharnhorst 1990).

- Bettencourt, L. M., D. I. Kaiser, and J. Kaur (2009). Scientific discovery and topological transitions in collaboration networks. *Journal of Informetrics* 3(3), 210–221. Special Edition on the Science of Science: Conceptualizations and Models of Science, ed. by Katy Börner & Andrea Scharnhorst.
- Bonitz, M. (1991). The impact of behavioral principles on the design of the system of scientific communication. *Scientometrics* 20(1), 107–111.
- Börner, K. (2010). *Atlas of Science: Visualizing What We Know*. MIT Press.
- Börner, K., J. T. Maru, and R. L. Goldstone (2004). The simultaneous evolution of author and paper networks. *Proceedings of the National Academy of Sciences of the United States of America* 101, 5266–5273.
- Börner, K., S. Sanyal, and A. Vespignani (2007). Network science. *Annual Review of Information Science and Technology* 41, 537–607.
- Börner, K. and A. Scharnhorst (2009). Visual conceptualizations and models of science. *Journal of Informetrics* 3(3), 161–172. Science of Science: Conceptualizations and Models of Science.
- Bradford, S. C. (1934). Sources of information on specific subjects. *Engineering* 137, 85–86.
- Brin, S. and L. Page (1998). The anatomy of a large-scale hypertextual Web search engine. *Computer Networks and ISDN Systems* 30(1–7), 107–117.
- Bruckner, E., W. Ebeling, and A. Scharnhorst (1990). The application of evolution models in scientometrics. *Scientometrics* 18(1), 21–41.
- Bruckner, E. and A. Scharnhorst (1986). Bemerkungen zum Problemkreis einer wissenschafts-wissenschaftlichen Interpretation der Fisher-Eigen-Gleichung dargestellt an Hand wissenschaftshistorischer Zeugnisse. *Wissenschaftliche Zeitschrift der Humboldt-Universität zu Berlin. Mathematisch-naturwissenschaftliche Reihe* 35(5), 481–493.
- Callon, M., J. P. Courtial, W. A. Turner, and S. Bauin (1983). From translations to problematic networks: An introduction to co-word analysis. *Social Science Information* 22(2), 191–235.
- Chen, C., Y. Chen, M. Horowitz, H. Hou, Z. Liu, and D. Pellegrino (2009). Towards an Explanatory and Computational Theory of Scientific Discovery. *Journal of Informetrics*. Special Edition on the Science of Science: Conceptualizations and Models of Science, ed. by Katy Börner & Andrea Scharnhorst.
- De Bellis, N. (2009). *Bibliometrics and Citation Analysis. From the Science Citation Index to Cybermetrics*. Lanham: The Scarecrow Press. ISBN-13: 978-0-8108-6713-0.
- de Solla Price, D. J. (1965). Networks of scientific papers. *Science* 149, 510–515.
- Deerwester, S., S. T. Dumais, G. W. Furnas, T. K. Landauer, and R. Harshman (1990). Indexing by latent semantic analysis. *Journal of the American Society for Information Science* 41(6), 391–407.
- Ebeling, W. and A. Scharnhorst (2009). Selbstorganisation und Mobilität von Wissenschaftlern – Modelle für die Dynamik von Problemfeldern und Wissenschaftsgebieten. In W. Ebeling and H. Parthey (Eds.), *Selbstorganisation in Wissenschaft und Technik. Wissenschaftsforschung Jahrbuch 2008*, pp. 9–27. Wissenschaftlicher Verlag Berlin.
- Edmonds, B., N. Gilbert, P. Ahrweiler, and A. Scharnhorst (2011). Simulating the social processes of science. *Journal of Artificial Societies and Social Simulation* 14(4), 1–6. <http://jasss.soc.surrey.ac.uk/14/4/14.html> (Special issue).
- Egghe, L. and R. Rousseau (1990). *Introduction to Informetrics. Quantitative Methods in Library and Information Science*. Elsevier.
- Fortunato, S. (2010). Community detection in graphs. *Physics Reports* 486(3–5), 75–174.
- Fronczak, P., A. Fronczak, and J. A. Hołyst (2007). Analysis of scientific productivity using maximum entropy principle and fluctuation-dissipation theorem. *Physical Review E* 75(2), 26103.
- Galam, S. (2011). Tailor based allocations for multiple authorship: a fractional gh-index. *Scientometrics* 89(1), 365–379.

- Garfield, E. (1955). Citation indexes for science. A new dimension in documentation through association of ideas. *Science* 122(3159), 108–111.
- Garfield, E. (1981). Introducing the ISI Atlas of Science: Biochemistry and Molecular Biology. *Current Contents* 42, 279–87. Reprinted: Essays of an Information Scientist, Vol. 5, p. 279–287, 1981–82 <http://www.garfield.library.upenn.edu/essays/v5p279y1981-82.pdf>.
- Garfield, E., A. I. Pudovkin, and V. S. Istomin (2003). Why do we need algorithmic historiography? *Journal of the American Society for Information Science and Technology* 54(5), 400–412.
- Garfield, E. and I. H. Sher (1963). New factors in the evaluation of scientific literature through citation indexing. *American Documentation* 14(3), 195–201. <http://www.garfield.library.upenn.edu/essays/v6p492y1983.pdf>, 2005-4-28.
- Geller, N. L. (1978). On the citation influence methodology of Pinski and Narin. *Information Processing & Management* 14(2), 93–95.
- Gilbert, N. (1997). A simulation of the structure of academic science.
- Glänzel, W. (2003). Bibliometrics as a research field: A course on theory and application of bibliometric indicators. Courses Hand-out, http://www.norslis.net/2004/Bib_Module_KUL.pdf.
- Gläser, J. (2006). *Wissenschaftliche Produktionsgemeinschaften: Die soziale Ordnung der Forschung*. Campus Verlag.
- Gross, P. L. K. and E. M. Gross (1927). College Libraries and Chemical Education. *Science* 66(1713), 385–389.
- Havemann, F. (2009). *Einführung in die Bibliometrie*. Berlin: Gesellschaft für Wissenschaftsforschung. <http://d-nb.info/993717780>.
- Havemann, F., J. Gläser, M. Heinz, and A. Struck (2012, March). Identifying overlapping and hierarchical thematic structures in networks of scholarly papers: A comparison of three approaches. *PLoS ONE* 7(3), e33255.
- Havemann, F. and A. Scharnhorst (2010). Bibliometrische Netzwerke. In C. Stegbauer and R. Häußling (Eds.), *Handbuch Netzwerkforschung*, pp. 799–823. VS Verlag für Sozialwissenschaften. 10.1007/978-3-531-92575-2_70.
- Hellsten, I., R. Lambiotte, A. Scharnhorst, and M. Ausloos (2007). Self-citations, co-authorships and keywords: A new approach to scientists’ field mobility? *Scientometrics* 72(3), 469–486.
- Huberman, B. A. (2001). *The laws of the Web: patterns in the ecology of information*. MIT Press.
- Janssens, F., W. Glänzel, and B. De Moor (2008). A hybrid mapping of information science. *Scientometrics* 75(3), 607–631.
- Kessler, M. M. (1963). Bibliographic coupling between scientific papers. *American Documentation* 14, 10–25.
- Klavans, R. and K. Boyack (2009). Toward a consensus map of science. *Technology* 60, 2.
- Kostoff, R. N. (2007). Literature-related discovery (LRD): Introduction and background. *Technological Forecasting & Social Change* 75(2), 165–185.
- Lambiotte, R. and P. Panzarasa (2009). Communities, knowledge creation, and information diffusion. *Journal of Informetrics* 3(3), 180–190.
- Latour, B. (2005). *Reassembling the social: An introduction to actor-network-theory*. Oxford University Press, USA.
- Laudel, G. (1999). Interdisziplinäre Forschungsk Kooperation: Erfolgsbedingungen der Institution ‘Sonderforschungsbereich’. Berlin, Edition Sigma.
- Laudel, G. (2002). What do we measure by co-authorships? *Research Evaluation* 11(1), 3–15.
- Leydesdorff, L. (2001). *The Challenge of Scientometrics: the development, measurement, and self-organization of scientific communications*. Upublish. Com.
- Leydesdorff, L. (2007). Visualization of the citation impact environments of scientific journals: An online mapping exercise. *Jour-*

- nal of the American Society for Information Science and Technology* 58(1), 25–38.
- Leydesdorff, L. and I. Hellsten (2006). Measuring the meaning of words in contexts: An automated analysis of controversies about ‘Monarch butterflies,’ ‘Frankenfoods,’ and ‘stem cells’. *Scientometrics* 67(2), 231–258.
- Leydesdorff, L. and T. Schank (2008). Dynamic animations of journal maps: Indicators of structural changes and interdisciplinary developments. *Journal of the American Society for Information Science and Technology* 59(11), 1810–1818.
- Lotka, A. J. (1926). The frequency distribution of scientific productivity. *Journal of the Washington Academy of Sciences* 16(12), 317–323.
- Lucio-Arias, D. and L. Leydesdorff (2008). Main-path analysis and path-dependent transitions in HistCite-based historiograms. *Journal of the American Society for Information Science and Technology* 59(12), 1948–1962.
- Lucio-Arias, D. and L. Leydesdorff (2009). The dynamics of exchanges and references among scientific texts, and the autopoiesis of discursive knowledge. *Journal of Informetrics*.
- Lucio-Arias, D. and A. Scharnhorst (2012). Mathematical approaches to modeling science from an algorithmic-historiography perspective. In A. Scharnhorst, K. Börner, and P. Besselaar (Eds.), *Models of Science Dynamics*, Understanding Complex Systems, pp. 23–66. Springer Berlin Heidelberg.
- Marshakova, I. V. (1973). System of document connections based on references. *Nauchno-Tekhnicheskaya Informatsiya Seriya 2 – Informatsionnye Protsessy i Sistemy* 6, 3–8. (in Russian).
- Merton, R. K. (1957). *Social theory and social structure*. Free Press New York.
- Milgram, S. (1967). The small world problem. *Psychology Today* 2(1), 60–67.
- Mitesser, O., M. Heinz, F. Havemann, and J. Gläser (2008). Measuring Diversity of Research by Extracting Latent Themes from Bipartite Networks of Papers and References. In H. Kretschmer and F. Havemann (Eds.), *Proceedings of WIS 2008: Fourth International Conference on Webometrics, Informetrics and Scientometrics Ninth COLLNET Meeting*, Berlin. Humboldt-Universität zu Berlin: Gesellschaft für Wissenschaftsforschung. ISBN: 978-3-934682-45-0, <http://www.collnet.de/Berlin-2008/MitesserWIS2008mdr.pdf>.
- Moed, H. F., W. Glänzel, and U. Schmoch (Eds.) (2004). *Handbook of Quantitative Science and Technology Research: The Use of Publication and Patent Statistics in Studies of S&T Systems*. Dordrecht: Kluwer Academic Publishers.
- Morris, S. A., G. Yen, Z. Wu, and B. Asnake (2003). Time line visualization of research fronts. *Journal of the American Society for Information Science and Technology* 54(5), 413–422.
- Morris, S. A. and G. G. Yen (2004). Crossmaps: Visualization of overlapping relationships in collections of journal papers. *Proceedings of the National Academy of Sciences of The United States of America* 101, 5291–5296.
- Mutschke, P., P. Mayr, P. Schaer, and Y. Sure (2011). Science models as value-added services for scholarly information systems. *Scientometrics* 89(1), 349–364.
- Newman, M. E. J. (2001a). Scientific collaboration networks. I. Network construction and fundamental results. *Physical Review E* 64, 016131.
- Newman, M. E. J. (2001b). Scientific collaboration networks. II. Shortest paths, weighted networks, and centrality. *Physical Review E* 64, 016132.
- Noyons, E. C. M. (2004). Science maps within a science policy context. In H. F. Moed, W. Glänzel, and U. Schmoch (Eds.), *Handbook of Quantitative Science and Technology Research: The Use of Publication and Patent Statistics in Studies of S&T Systems*, pp. 237–256. Dordrecht: Kluwer Academic Publishers.
- NRC (2005). Network science. National Research Council of the National Academies, Washington, DC.

- Ortega, J. L., I. Aguillo, V. Cothey, and A. Scharnhorst (2008). Maps of the academic web in the European Higher Education Area—an exploration of visual web indicators. *Scientometrics* 74(2), 295–308.
- Payette, N. (2012). Agent-based models of science. In A. Scharnhorst, K. Börner, and P. Besselaar (Eds.), *Models of Science Dynamics*, Understanding Complex Systems, pp. 127–157. Springer Berlin Heidelberg.
- Pinski, G. and F. Narin (1976). Citation influence for journal aggregates of scientific publications: theory, with application to literature of physics. *Information Processing & Management* 12, 297–312.
- Prabowo, R., M. Thelwall, I. Hellsten, and A. Scharnhorst (2008). Evolving debates in online communication – a graph analytical approach. *Internet Research* 18(5), 520–540.
- Pyka, A. and A. Scharnhorst (2009). Introduction. Network perspectives on innovations: Innovative networks – network innovation. In A. Pyka and A. Scharnhorst (Eds.), *Innovation Networks – New Approaches in Modelling and Analyzing*. Berlin: Springer.
- Rip, A. and J. Courtial (1984). Co-word maps of biotechnology: An example of cognitive scientometrics. *Scientometrics* 6(6), 381–400.
- Salton, G. and M. J. McGill (1983). *Introduction to Modern Information Retrieval*. New York, NY, USA: McGraw-Hill, Inc.
- Scharnhorst, A. (2001). Constructing Knowledge Landscapes within the Framework of Geometrically Oriented Evolutionary Theories. In M. Matthies, H. Malchow, and J. Kriz (Eds.), *Integrative Systems Approaches to Natural and Social Sciences – Systems Science 2000*, pp. 505–515. Berlin: Springer.
- Scharnhorst, A. (2003, July). Complex networks and the web: Insights from nonlinear physics. *Journal of Computer-Mediated Communication* 8(4). <http://jcmc.indiana.edu/vol8/issue4/scharnhorst.html>.
- Skupin, A. (2009). Discrete and continuous conceptualizations of science: Implications for knowledge domain visualization. *Journal of Informetrics* 3(3), 233–245. Special Edition on the Science of Science: Conceptualizations and Models of Science, ed. by Katy Börner & Andrea Scharnhorst.
- Small, H. (1973). Co-citation in the Scientific Literature: A New Measure of the Relationship Between Two Documents. *Journal of the American Society for Information Science* 24, 265–269.
- Small, H. (1978). Cited documents as concept symbols. *Social studies of science* 8(3), 327.
- Small, H. and E. Sweeney (1985). Clustering the Science Citation index using co-citations. *Scientometrics* 7(3), 391–409.
- Snijders, T. A. B., G. G. van de Bunt, and C. E. G. Steglich (2010). Introduction to stochastic actor-based models for network dynamics. *Social Networks* 32(1), 44–60.
- Sonnenwald, D. (2007). Scientific collaboration. *Annual Review of Information Science and Technology* 41(1).
- Swanson, D. R. (1986). Fish oil, Raynauds syndrome, and undiscovered public knowledge. *Perspectives in Biology and Medicine* 30(1), 7–18.
- Thelwall, M. (2004). *Link analysis: An information science approach*. Academic Press.
- Thelwall, M. (2009). Introduction to Webometrics: Quantitative Web Research for the Social Sciences. In *Synthesis Lectures on Information Concepts, Retrieval, and Services*, Volume 1, pp. 1–116. Morgan & Claypool Publishers.
- van den Heuvel, C. (2008). Building Society, Constructing Knowledge, Weaving the Web. Otlet’s visualizations of a global information society and his concept of a universal civilization. In W. B. Rayward (Ed.), *European Modernism and the Information Society*, Chapter 7, pp. 127–153. London: Ashgate Publishers.
- van den Heuvel, C. and W. Rayward (2011). Facing interfaces: Paul Otlet’s visualizations of data integration. *Journal of the American Society for Information Science and Technology* 62(12), 2313–2326.

- van den Heuvel, C. and W. B. Rayward (2005, July). Visualizing the Organization and Dissemination of Knowledge: Paul Otlet's Sketches in the Mundaneum, Mons', Envisioning a Path to the Future. Weblog. http://informationvisualization.typepad.com/sigvis/2005/07/visualizations_.html.
- van der Eijk, C. C., E. M. van Mulligen, J. A. Kors, B. Mons, and J. van den Berg (2004). Constructing an Associative Concept Space for Literature-Based Discovery. *Journal of the American Society for Information Science and Technology* 55(5), 436–444.
- Vlachý, J. (1978). Field mobility in Czech physics-related institutes and faculties. *Czechoslovak Journal of Physics* 28(2), 237–240.
- Wagner, C. S. (2008). The new invisible college. Science for Development, Brookings Institution, Washington, DC.
- Wasserman, S. and K. Faust (1994). *Social Network Analysis: Methods and Applications*. Cambridge University Press.
- Wouters, P. (1999). The citation culture. *Unpublished Ph. D. Thesis, University of Amsterdam*. <http://garfield.library.upenn.edu/wouters/wouters.pdf>.